Can Games Be AI Explanations? An Exploratory Study of Simulation Games

Jennifer Villareale[1], Thomas Boyd Fox[2], and Jichen Zhu[3]
[1]Drexel University , Philadelphia, PA, USA, jmv85@drexel.edu
[2]Drexel University, Philadelphia, PA, USA, thomas.boyd.fox@gmail.com
[3]IT University of Copenhagen, Copenhagen, Denmark ,
jichen.zhu@gmail.com

ABSTRACT

This paper explores the potential of computer games as a new form of explainable interface to AI. Most existing eXplainable AI (XAI) provide explanations in static or limited interactive forms. This paper analyzes simulation games as an exploratory case study based on an established taxonomy in human-centered XAI. Our analysis indicates that the existing mechanics and game interfaces of simulation games, to various extents, can support most XAI question types, although certain XAI question types are difficult to convey. We offer initial reflections on the design space of leveraging insights from games to rethink the explainable interfaces for XAI.

Keywords

eXplainable AI, computer games, game design

INTRODUCTION

Simulation games are a category of computer games that emulate the functioning of complex systems, which consist of many interacting components that produce emergent behaviors and outcomes. Complex systems, such as ecology, economy, society, and biology, are often difficult to understand and manage. By simulating complex systems, simulation games can offer players the opportunity to explore, experiment, and experience the system dynamics in a virtual environment. Research has shown that simulation games can help players develop a deeper understanding of the system structure, behavior, and consequences, as well as foster systems thinking, problem-solving, and critical thinking skills (Sitzmann 2011; Squire et al. 2012; Vlachopoulos et al. 2017). Games researchers have used both commercially available simulation games (e.g., *Civilization* series, which models historical and cultural development (Squire et al. 2012)) and custom-designed games (e.g., *Parallel*, which models concurrent and parallel programming (Zhu et al. 2019)) to facilitate learning.

This paper explores a new approach to using *simulation games as a format of AI explanation.* As AI becomes widely adopted in all areas of society, there is an urgent need to improve these systems' fairness, accountability, and transparency. The rapidly growing research area of eXplainable AI (XAI) is one of the most promising approaches towards these goals by opening the AI black box and explaining its

Proceedings of DiGRA 2024

©2023 Authors & Digital Games Research Association DiGRA. Personal and educational classroom use of this paper is allowed, commercial use requires specific permission from the author.

underlying operation to humans (Gunning 2017). While XAI researchers have made significant progress in increasing AI's *explanability* (Ribeiro et al. 2016; Shrikumar et al. 2017), most *explanations* produced by XAI currently lack usability, practical interpretability, and efficacy for real users (Abdul et al. 2018; Doshi-Velez et al. 2017; Zhu et al. 2018; Zhu et al. 2020). A key knowledge gap is how to explain the complex operations of a machine learning model in ways humans can understand — that is, to turn technical *explainability* into user-centered *explanations* (Nguyen et al. 2024).

In this paper, we explore the feasibility of simulation games as a new form of interactive explanations. We do so by examining existing mechanics in simulation games and identifying how simulation games communicate the inner workings of complex models to players, especially through gameplay and interfaces. Specifically, we use an established taxonomy of XAI explanations developed by Liao et al. (2021) as our foundation and identify salient examples of how existing simulation games convey similar types of information. The goal of the paper is not to provide a comprehensive survey of simulation games but to present an initial exploration of whether simulation games can be the basis of a potential format of AI explanations.

In the rest of the paper, we first discuss related work and then present our analysis on whether and how existing mechanics of simulation games can provide key types of information required by XAI. Finally, we discuss the initial steps towards designing simulation games as a new type of explainable interface for XAI.

RELATED WORK

This section discusses related work on XAI, explanation interfaces, and existing research on using games for XAI.

Explainable AI (XAI) and XAI Question Bank

The rapidly growing field of XAI has made significant breakthroughs in *technical explainability*, producing established XAI algorithms such as LIME (Ribeiro et al. 2016), DeepLIFT (Shrikumar et al. 2017), and LRP (Binder et al. 2016). Typically, XAI explanations either reveal the inner workings of an ML model or demonstrate the models' reasoning in a post-hoc way. The first technique suits more human-understandable models (e.g., rule-based), whereas the second works better for less interpretable models (e.g., deep neural networks). Technical reviews of XAI can be found in (Adadi et al. 2018; Arrieta et al. 2020).

This paper draws on the human-centered XAI framework developed by Liao et al. (2020) and Liao et al. (2021). Through interviews with XAI researchers, Liao and colleagues proposed an *XAI Question Bank*, in which user needs for explainability are represented as prototypical questions users might ask about the AI. Broadly, they propose a series of nine questions that encapsulate what exactly different XAI techniques seek to explain: 1) How (global model-wide), 2) Why (a given prediction), 3) Why Not (a different prediction), 4) How to Be that (a different prediction), 5) How to Still Be This (the current prediction), 6) What if, 7) Performance, 8) Data, and 9) Output. While Liao et al. state that the above list is not exhaustive, their taxonomy is a useful starting point for analyzing how simulation games communicate similar types of information to players.

Explainable Interface

Despite the technical advancements in XAI, most explanations produced by XAI lack usability, practical interpretability, and efficacy for real users (Abdul et al. 2018; Doshi-Velez et al. 2017; Miller 2019; Zhu et al. 2018). A recent study found that a significant group of users (over 30%) could not understand the XAI explanations sufficiently well to use them even in relatively simple tasks (Narayanan et al. 2018). A key knowledge gap is how to explain the vastly complex operations of a machine learning model in ways humans can understand — that is, to turn technical *explainability* into user-centered *explanations*.

To bridge this gap, a growing body of XAI research has focused on how to design *explainable interfaces* (Chromik et al. 2021; Nguyen et al. 2024; Haque et al. 2023), also known as explanation format and user experience of XAI explanations (Liao et al. 2020; Liao et al. 2021), in ways that are useful to actual users, especially non-technical ones (Ghajargar et al. 2022). An explainable interface is the user interface through which human users access the explanations generated by XAI algorithms (Chromik et al. 2021; Mohseni et al. 2021; Mueller et al. 2019). Note that the *content* of the explanation, generated by an XAI algorithm such as LIME (Ribeiro et al. 2016), can be represented in different explainable interfaces. For instance, an explainable interface can be a paragraph of text explaining why an ML model denied an applicant's loan application and which key features (e.g., income level, education, gender) contributed to the ML decision. It could also be an interactive table that allows users to play with different values of the key features and see how that affects the ML model's decision (Cheng et al. 2019).

Among explainable interface design research, how to incorporate interactivity is a critical open problem (Chromik et al. 2021). With some exceptions, notably in visualization-based XAI (Liu et al. 2017; Choo et al. 2018; Spinner et al. 2019), most explainable interfaces are in static or limited interactive forms (Abdul et al. 2018). This paper looks into how simulation games convey similar information and explanations to draw inspiration for the design of explainable interfaces.

Games and XAI

A relatively small but growing body of research has looked into how computer games can help people understand AI (Myers et al. 2020; Fulton et al. 2020; Pemberton et al. 2019; Zhu et al. 2018; Villareale et al. 2021) and provide different types of player-AI interaction (Zhu et al. 2021). More recently, some games research directly explores the intersection of games and XAI (Zhu et al. 2018; Fulton et al. 2020; Sevastjanova et al. 2021; Xie et al. 2019). For example, Fulton et al. (2020) developed *XAI for Image Recognition*, a multiplayer explanation game designed to assess how humans interpret AI explanations for a deep learning model trained for image recognition. They found that games offered a valuable domain to understand how humans select and interpret explanations. Sevastjanova et al. (2021) developed *QuestComb*, a game to support complex classification tasks for a machine learning model, such as training and optimization. They found that gameplay helped engage users to create effective training data, enabled by continuous feedback from

the game environment. Xie et al. (2019) explored how *QUBE*, a simulation-based game, impacts designers' understanding of Machine Learning (ML) concepts. Results showed that designers' high-level ML understanding significantly increased after using *QUBE*. This paper extends these game design-based projects by exploring simulation games as a genre of computer games and their feasibility in providing players with the information required for XAI.

MAPPING XAI QUESTIONS TO SIMULATION GAME DESIGN

In this section, we present an exploratory study on whether existing mechanics and interfaces of simulation games *are able to* provide the type of information needed for XAI, identified in the XAI Question Bank (Liao et al. 2021). As an initial *affordance* study, our focus is to assess which types of key XAI information are already delivered in some well-known simulation games, which types are not yet supported by possible, and which are not compatible with simulation games. Since few simulation games are designed to emulate ML models, our analysis includes games that simulate other complex systems. We acknowledge that there are multiple differences between, for example, a facial recognition deep neural network and a farming simulator. It is possible that some of the design elements we identified are not applicable to ML-specific simulation games. Future research should further explore how to design simulation games specifically for ML models.

Methods

We first adapted Liao et al. (2020) *XAI Question Bank* from the domain of UX design to simulation games. Two researchers discussed and developed an adapted definition for each of the six question categories (i.e., *How, Why, Why Not, How To Be That, How To Still Be This*, and *What If*). In this preliminary analysis, we excluded the *Performance, Data, Output,* and *Other* categories due to their generic nature. The researchers then iteratively tested and refined the definitions by applying them to simulation game examples. Collectively, two researchers played and watched gameplay videos of the games we analyzed. Table 1 provides an overview of the six XAI question categories we analyzed, our adapted definition, and the associated game examples. For simplicity, we include Liao et al.'s original definitions in the text below.

Analysis

Below, we discuss examples of whether and how information about the six main question categories are included in some existing simulation games.

"How" Questions

As one of the most important questions that XAI attempts to explain, the "How" question provides information on how a model works at a global model-wide level. Typical ways to explain include "describe the general model logic as feature impact, rules, or decision trees," or "if a user is only interested in a high-level view, describe what are the top features or rules considered" (Liao et al. 2020).

In our analysis, we find that simulation games are particularly well-suited to communicate a model's rules and operations through experiencing and participating in

XAI Question Type	Adapted Definition	Examples from Simulation Games
How	How does the underlying system make decisions?	*Farm Simulator 19, Sid Meier's Civilization, Democracy 3, RimWorld, Farm Simulator 19, Roller Coaster Tycoon, Zoo Tycoon, Dwarf Fortress*
Why	Why/how is this particular outcome in the game reached?	*The Sims 4, Crusader Kings, Farm Simulator 19, Stellaris, Crusader Kings*
Why Not	Why/how is this outcome NOT reached?	*Dr. Derks Mutant Battleground*
How To Be That	How could this outcome be changed to that one?	*Farm Simulator 19, How To Train Your Snake*
How To Still Be This	What is the scope of change permitted to still get the same outcome?	*Poly Bridge, Poly Bridge 2*
What If	What would the system decide if this situation changes to...?	*The Sims 4, Democracy 3, Kerbal Space Program*

Figure 1: Adapted XAI question taxonomy to Simulation games.

the core gameplay loop. Note that hereafter, we broadly use the term "model" to refer to the games' underlying complex system, not just machine learning models. The game in this category (i.e., Farming Simulator) addresses the "How" question by allowing players to 1) change core aspects of the complex system (e.g., farming, managing economies) through various game mechanics (e.g., planting, spending money) and 2) observe the impact of their in-game actions and decisions on the overall operation of the model in terms of its success in the game world. We consider this an implicit way to address the question, as information does not prompt the player or is displayed for players to review. Here, answers to "how the model works" are intended to be discovered over time.

Farming Simulator 19 (GIANTS Software 2018), a farming simulation game that allows players to grow a farm's economy by performing various tasks. To succeed in the game, players must prepare land for planting, tend to crops, raise livestock, and harvest crops, all while managing the farm's finances. The game addresses the "How" question of *how does the economy of a farm operate*. It does so by allowing players to manipulate the core farming elements (i.e., tractor equipment, tilling the soil, watering, and fertilizing) that influence the farm's success in terms of its production. Here, players discover the general logic and rules of managing a farm over time through actively participating in and experiencing the core gameplay loop. For example, players learn the general rules of each farming season regarding which crops are appropriate to plant and harvest and which weather conditions to be aware of through trial and error gameplay and actively engaging with the game's seasons menu (see Figure 2). The menu overviews when players should plant and harvest their crops. Accessing the other various tabs in the menu also provides players with information on the season's weather forecast and how each crop may be impacted by potential frost or drought to support players in understanding the rules and conditions of each crop and what the farm needs for the current or future

seasons.

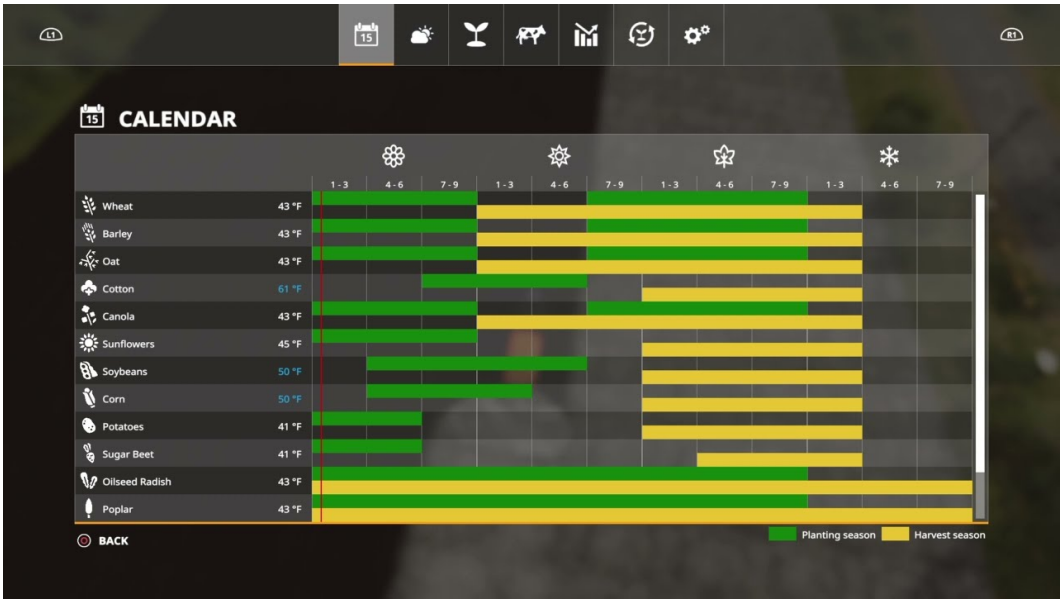

Figure 2: The Seasons Menu in *Farming Simulator 19*, which allows players to view different seasons of the year regarding when to plant and harvest their crops so players can adapt to changing conditions.

"Why" Questions

In Liao et al.'s framework, the "Why" question aims to explain why a particular model prediction, instance, or outcome has occurred. Typical ways to explain include describing "how features of the instance, or what key features, determine the model's prediction of it," "rules that the instance fits to guarantee the prediction," or showing examples with the "same predicted outcome to justify the model's prediction" (Liao et al. 2020). The key difference between the "how" and "why" questions are the former focuses on how the overall system works while the latter focuses on explaining why a particular outcome is reached.

The games in this category address the "Why" question by providing players with visualizations (e.g., colored bar charts) or text-based information to examine when making in-game decisions. Players iteratively refer to or seek out this information to gain more context on their progress. We consider this an explicit way to address the question, as information can be accessed in alternate menus or the game user interface to help players understand why the model's current or future state may or may not be successful.

The Sims 4 (Maxis 2014) is a key example of a game that uses simple visualizations to help players understand the model's current state and make in-the-moment decisions using the game UI. In the game, players manage a virtual Sim character by choreographing their daily routine and life choices (e.g., career, marriage) while keeping their Sim characters alive and happy. To be successful, players iteratively refer to the "needs menu" (see Fig 3), a colored bar chart that turns from green to

orange to red depending on the severity of the Sim's need. For example, if the Sim's hygiene bar is red, the Sim requires a shower or bath to regain the Sim's hygiene status. Overall, this menu answers "Why" their Sim is currently unhappy and what features determine this state since a low score in any of these features (e.g., bladder, hygiene, hunger, social) will lead to an unhappy Sim. This menu allows players to make a more informed decision.

Figure 3: The Needs Menu in *The Sims 4*, which helps players answer why their Sim is currently unhappy and what features determine the current state of the Sim.

Alternatively, the game *Crusader Kings 3* (Paradox Interactive 2020) offers text-based information to help players understand the future state of the model if they were to maintain their current course of action. In the game, players control and manage a character and their dynasty, alongside their political, economic, and military issues. During gameplay, players can view a predictive battle screen (see Figure 4) that predicts the future state of the player character's exploits. For example, the battle screen indicates positive features (e.g., Better Army Commander, More Commander Traits, More Soldiers) in green and negative features (e.g., Defending in Wetlands, Higher Quality) in red to provide context to the outcome (e.g., win, lose), and to provide players with a direction to go in if they are not happy with the predicted outcome.

"Why Not" Questions

The "Why Not" question aims to justify why a different model prediction or outcome has *not* occurred. Typical ways to explain include describing "what features of the instance determine the current prediction and with what changes the instance would get the alternative prediction" or showing prototypical examples of "alternative outcomes."

Figure 4: The predictive battle screen in **Crusader Kings 3**, which provides predictions on how upcoming battles may go and what features might change the outcome.

We found one game, *Dr. Derks Mutant Battleground* (Mount Rouke Studios 2020) implicitly answered the "Why Not" question by providing players with supportive animations showing alternative outcomes. In the game, players train mutants called Derklings, each with their own Neural Network, to learn offensive and defensive tasks important for battle. To train derklings, players iteratively set rewards and punishments and "run" the training session, which provides players with an animation (see Figure 5) of the entire "population" of Derklings in each generation training toward the goal the player set. For example, training the Derkling by setting rewards for hitting an opponent's tower. In this case, answering the "Why Not" question becomes important to help players understand which alternative Derkling could be chosen to succeed the next generation and assess overall how the training is going. During the animation, players can view alternative outcomes by clicking on each Derkling within the same population. Then, players may decide which may succeed in the next generation and adjust their gameplay to accommodate the outcome.

"How To Be That" Questions

The "How To Be That" question aims to explain how to obtain a different model prediction or outcome. Typical ways to explain include "highlighting feature(s) that if changed (increased, decreased, absent, or present) could alter the prediction to the alternative outcome, with minimum effort required" or showing examples with "minimum differences but had the alternative outcome."

The games in this category explicitly address the "How To Be That" question by providing players with information on which features need to be changed to obtain a different or more successful outcome through supplementary menus or implicitly through supportive animations or visuals. Answering the "How To Be That" question becomes particularly important when players develop or refine their strategy. Here, examples are not shown to the player. Instead, the focus is on discovering and building strategies by utilizing these menus and animations, which aim to highlight features to be changed to support players in tailoring their gameplay.

A good example of a game explicitly providing "How To Be That" information is *Farming Simulator 19*. In the game, players can access a map menu that contains a

Figure 5: A screenshot from the training animation in *Dr. Derks Mutant Battleground*. The small triangle shapes represent other Derklings in the population training to attack an opposing player's tower.

filter system in which the player can select various filters to overlay across the whole map to examine different information. For example, the overlay shows the player which part of their owned land needs attention and how it can be changed toward a better outcome, such as "needs lime" or "needs plowing." This provides players with specific information on how to improve their farm production. It also highlights if certain features (e.g., temperature, fertilization) have increased or decreased to support players in understanding how to change their strategy to improve their farm.

How To Train Your Snake (bewelge 2017), on the other hand, implicitly provides "How To Be That" information through animations that showcase the model's progress. In the game, modeled after the original game *Snake*, players must steer each snake that a Neural Network controls by selecting various upgrades to make the snakes reach a particular length by eating food. The game visualizes the snake's progress by showing each snake navigating the game board to find the food (see Figure 6). By observing the snake's progress and assessing its success based on this animation, players address "How To Be That" questions and decide which parameters need to be upgraded next (e.g., "Double Food Value").

"How To Still Be This" Questions

The "How To Still Be This" question aims to explain how to maintain a current model prediction, instance, or outcome. Typical ways to explain include describing "features/feature ranges or rules that could guarantee the same prediction " or showing examples that are "different from the instance but still had the same outcome."

Within the context of games, we found that this question was answered to allow users to optimize or experiment with their approaches to problems within the game

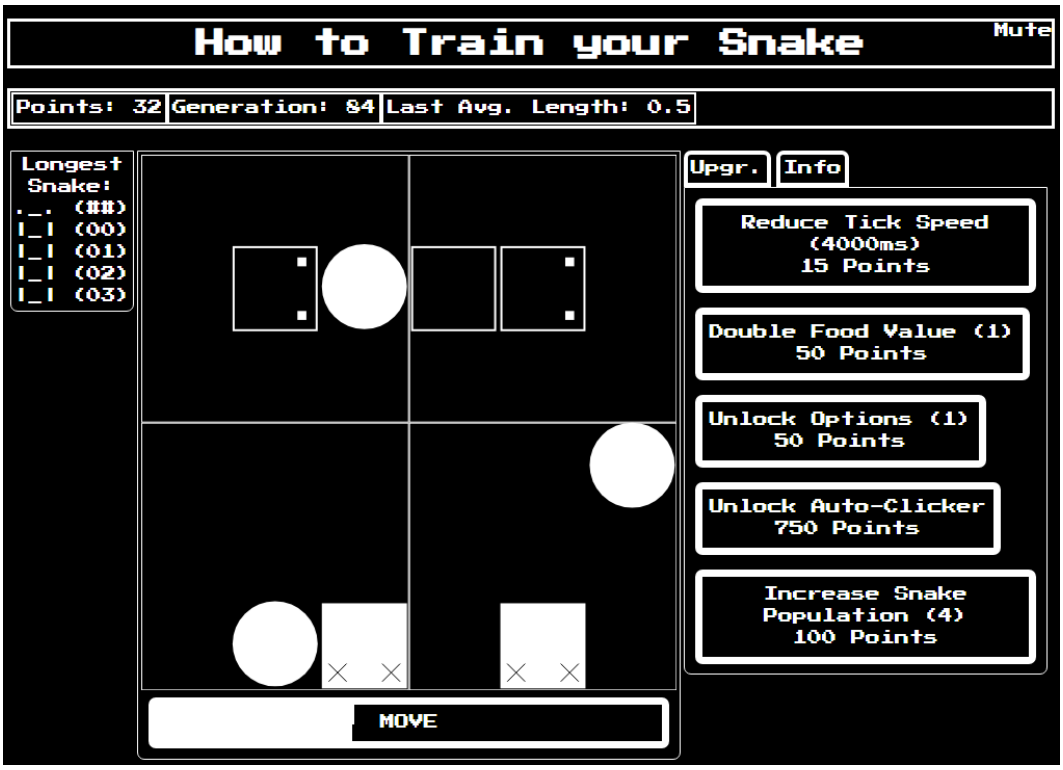

Figure 6: An example progress animation in *How To Train Your Snake*. The snake (i.e., the square in each quadrant), navigates the board to find food (i.e., circle) which causes it to grow in length.

model. By being presented with similar but distinct solutions to tasks within a game, the player can begin to engage with the model in different ways than they were initially. We consider this an explicit answer to the "How To Still Be This" question as the user is intentionally shown this information for the purpose of exploring the game's model.

For an example of this question in practice, we look to *Poly Bridge* (Dry Cactus 2016). *Poly Bridge* is a physics-simulation puzzle game where users are tasked with building bridges to allow various vehicles to cross a river. After completing a level, users are presented with data on how their solution performed in three categories: joint stress (the closest any part of the bridge came to breaking), cost (the price of the bridge), and material footprint (how much material, in meters, was used) with all stats presented both empirically and relative to the community. The sequel, *Poly Bridge 2* (Dry Cactus 2020), expands this to also allow users to immediately see the solutions of other players via the Gallery. This can allow the player to experiment with alternative strategies or goals, such as optimizing for a cheaper bridge or trying less intuitive bridge designs they might not have considered. Afterward, the user can take what they learned from this exploration into later, more demanding levels.

Figure 7: The level completion screen in *Poly Bridge 2*. The player is able to view the solutions of other players on the left side of the screen, replay their own solution in the center, and see how they compare to the global community on the right side.

"What If" Questions

The “What If” question is another core question that XAI aims to explain how a model outcome could be changed through the manipulation of its input. Typical ways to explain include showing examples of “how the prediction changes corresponding to the inquired change of input.”

In our analysis, we find that simulation games are particularly well-suited to communicate how the model can change corresponding to its change of input as this inherently is what gameplay is, manipulating inputs (i.e., actions in a game) and receiving an output (i.e., feedback that the game provides) creating an innate interaction loop that naturally encourages “What If” explorations. Similar to the “How” category, the games in this category address the “What If” question by allowing players to 1) change core aspects of the complex system (e.g., farming, managing economies) through various game mechanics (e.g., planting, spending money) and 2) observe the impact of their in-game actions and decisions on the overall operation of the model. The difference between this category and the “How” category is the player-provided goal to explore “What If” through exploring different hypotheses. Here, games address“What If” by providing the mechanics necessary to explore different input scenarios.

While all the games in this paper are examples of implicitly providing answers to“What If” questions, one game, specifically *Democracy 3* (Positech Games 2013), uniquely uses visualizations to support this question, showing examples of how an in-game decision from the player would change the state of the model. In the game, players manage a country’s policies by acting as its President. The player can introduce, remove, or adjust policies and explore the effects these will have on several socioeconomic factors and voter perception. When making changes to a policy, the game

provides players with a node-link diagram (see Fig 8) showing what other issues are directly linked to that policy and the immediate effects on voters' perception through colored bars when expanded. "What If" becomes particularly important as players simulate potential policy and budget changes before committing to the change, as it is relatively easy to lose the game if these predictions are not utilized effectively.

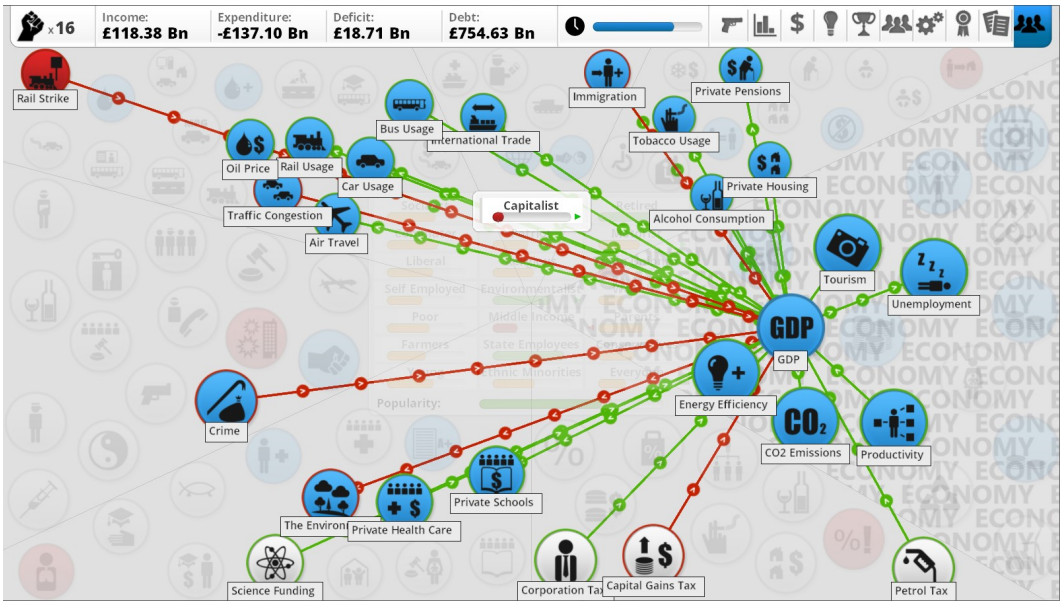

Figure 8: The policy visualization in *Democracy 3*

.

DISCUSSION: CAN GAMES BE AI EXPLANATIONS?

In this paper, we map different game examples to the key types of XAI questions. We offer a few reflections on the design space for using games as a new form of explanation.

Reflection 1: Explaining "What If" and "How" Through Play. Games offer players agency by giving players the ability to make meaningful choices that affect the outcome of the game, which arguably is core to what makes games engaging and fun (Koster 2013; Hodent 2020). In the case of simulation games described in this paper, players are able to make decisions in the game that have a significant impact on the game world. Here, each "game world" represents a model (e.g., farming, managing economies) that players are trying to decipher by tinkering with aspects of the model (i.e., planting, allocating resources) and observing the impact of their change, which is essential for learning and reflection (Juul 2013; Gee 2003). Interestingly, experiencing and participating in gameplay address the "What If" and "How" categories through play.

For example, in *Sim 4*, players tinker with various routines (e.g., cooking, cleaning) and lifestyle choices (i.e., career, marriage) to observe their impact on the life of their Sim character. Exploring "What If" questions contributes to not only the fun of exploring different scenarios but also provides a better understanding of how these

choices might impact the overall life of a Sim. In *Farming Simulator 19*, players explore different combinations of equipment, crops, and livestock to maximize their farm's production. Exploring "What If" scenarios using the game's supplementary menus (e.g., season menu, map filtering system) allows players to refine their approach and better understand what it takes to manage a farm. The key takeaway is that the experience of playing with "What If" questions explains "How" the operation of the model works in the game world over time.

Reflection 2: Why "Why Not" is Difficult. From our review, we found the question "Why Not" posed a challenge in finding clear, explicit examples in simulation games where this question had been addressed.

We found it difficult to find game examples that provided players with justifications on why a different model prediction or outcome has *not* occurred, perhaps due to the emphasis on games encouraging experimentation, which lends itself better to "What If" questions. For example, players are able to explore how adjusting different policies might affect several socio-economic factors and voter perception in *Democracy 3*. Through these experiences, players could develop an intuition for why something did *not* occur in the game; however, this remains implicit and player-driven. Another contributing factor could be that the "Why" question is more readily applicable within the context of games, with "Why Not" ending up redundant for many use cases. We found examples such as in *Crusader Kings 3* where users are presented an explanation for the outcome of a specific event. In scenarios where the outcome is (mostly) binary, the two questions largely become the same, as asking "Why" about a failure ends up being very similar in practice to asking "Why Not" about a success. One could imagine an interface where a player could specify a counterexample and then get information on factors that would cause or prevent such an outcome, but we did not find any examples in this vein. As we noted earlier, the visualization in *Dr. Derks Mutant Battleground* addressed by displaying alternative outcomes but stopped short of explaining them. Ultimately, we suggest XAI researchers further explore how games can be used to show prototypical examples and allow players to experience these first-hand.

Reflection 3: The Struggle of Implicit and Explicit Explanations During Gameplay. Simulation games address XAI questions both explicitly and implicitly and with varying degrees of success. For example, some games only partially address some questions' definitions (i.e., "How To Still Be This," "Why Not") while others did well to capture the entire question (i.e., "How," "What If"). We found a useful way to look at each game is to examine if they implicitly (through gameplay) or explicitly (through visualizations and text) addressed each category. For example, games that addressed "How" and "What If" questions allowed players to discover this information independently through play. On the other hand, some games explicitly provided content to answer questions, specifically "Why" "How To Be That" through visualizations and text-based information in menus. However, there remain a few considerations.

In particular, when players are directly tinkering with the model in these games, it

is not always clear what these mechanics do or what they mean concerning the model. For example, in *Crusader Kings 3*, a player might want to address some of the factors listed in the battle menu (see Figure 4) to improve the prediction. Therefore, inexperienced players may struggle to understand what these factors mean and how to change the system (e.g., Better Army Commander). Most of these questions can be answered in the game, but they require the player to fail frequently, make inferences, and learn through play. Here, presenting the correct answer on the screen would negate the fun of discovering this over time.

Further, it is well established that user behavior is more permissive in games (Frazier et al. 2012; Williams 2010) and provides freedom to explore different questions and hypotheses. Because of this, it may be difficult for researchers to guide players toward a particular explanation or outcome without removing the feeling of agency because players may experience the model differently depending on gameplay. This can be problematic when trying to explain a model. For example, *Crusader Kings 3* offers different paths to achieve the objective, such as choosing between a stealthy approach or a more direct, combat-focused approach. Depending on which path is taken, this may contribute to different understandings of the underlying model.

Towards Playful Explanation Interfaces

Recently, XAI researchers have been exploring the interaction afforded in explanation interfaces (Chromik et al. 2021; Cheng et al. 2019; Haque et al. 2023) and how these interfaces can help people build up an understanding of AI and its behavior through interactivity. For instance, Cheng et al. (2019) presents an interface that allows users to observe how the predictions of a university admission classifier change by allowing them to adjust the values of input features of applicants freely. This exploratory approach has sparked interest in the XAI community (Abdul et al. 2018; Chromik et al. 2021). However, keeping users engaged in these interfaces to ensure they gain the benefits of these interfaces, as extended interaction has shown to improve user comprehension of the system (Cheng et al. 2019), remains an area for improvement. Below, we offer some initial guidance on improving explanation interfaces with simulation game elements observed from our analysis.

Design Consideration 1: Provide objectives to direct users attention to important aspects of the system. We found that many games provided the player with objectives by visualizing the model's state. These visualizations served to direct the players' focus to specific parts of the model for them to iterate on. For example, *The Sims* provides the user with the "Needs Menu," which directs players' attention to different aspects of the system, thus prompting them to explore different approaches to meet those needs, such as cooking different meals or ordering take-out to improve the Sim's hunger bar. This provides a clear direction on what part of the model needs to change and what aspects users are able to manipulate. We suggest adding objectives by visualizing the state of the model in terms of which aspects are controlled by the user as a natural way to offer direction and focus the user's exploration of the system.

Design Consideration 2: Use failure to engage users' interest in exploring

the bounds of the system. In our analysis, we found that some games engaged interaction with the model by starting the model in a failed state or highlighting the possibility of the model's failure to players. For example, in the game *Dr. Derks Mutant Battleground* players started the game with Derklings, who could not perform any combative tasks and ran off the edges of the map. This engaged users to improve the model, thus gaining experience in the bounds of the system in terms of different failed states and the conditions for improving it. In contrast, *Crusader Kings* highlights the possibility of failure (see Fig 4) and the contributing factors for the user. This game highlighted the potential for failure in the UI, allowing gameplay to center on finding different ways to get the model to perform better or in a different way. We suggest exploring the role of playful failure as this could engage users to explore the bounds of the system and gain experience on how to improve the state of the system and what inputs are needed to do so.

Design Consideration 3: Use visualizations to support "What If" and "How" moments. We found that the simulation games were particularly good at explaining "What If" and "How" implicitly through play. However, these games always supported another secondary question, such as "Why" and "How To Still Be This" through visualizations to support the gameplay. In other words, these games supported player exploration with visualizations that explicitly answered a secondary question that players would need answered while playing with the model over time. For example, in the game *Farming Simulator*, the players can access a Seasons menu that contains information about when to harvest crops and what seasonal conditions to prepare for. This provides players with specific information on how to improve their farm production and further supports the "What If" moments. We suggest researchers focus interactions on allowing users to explore"What If" and "How" questions on the model, but provide explicit explanations on supportive questions important to the interaction, such as "How To Be That."

CONCLUSION

This paper proposes that computer games can be a new form of explainable interface and argues that games offer a rich set of interactions when communicating the inner workings of complex models to non-technical users. In this paper, we utilize Liao et al. (2020) *XAI Question Bank* and identify salient examples of how existing simulation games convey similar types of information and analyze their design. We offer reflections on the design space of leveraging insights from simulation games to rethink the explainable interfaces for XAI.

BIBLIOGRAPHY

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y. and Kankanhalli, M. 2018. "Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda." In *Proceedings of the 2018 CHI conference on human factors in computing systems,* 1–18.

Adadi, A. and Berrada, M. 2018. "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)." *IEEE access* 6:52138–52160.

Arrieta, A. B., Dı́az-Rodŕıguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcı́a, S., Gil-López, S., Molina, D., Benjamins, R. et al. 2020. “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI.” *Information fusion* 58:82–115.

bewelge. 2017. *How To Train Your Snake.* PC Game, Digital. bewelge.

Binder, A., Montavon, G., Lapuschkin, S., Müller, K.-R. and Samek, W. 2016. “Layer-Wise Relevance Propagation for Neural Networks with Local Renormalization Layers.” In *Artificial Neural Networks and Machine Learning,* edited by V. A., M. P. and P. R. A.

Cheng, H.-F., Wang, R., Zhang, Z., O’Connell, F., Gray, T., Harper, F. M. and Zhu, H. 2019. “Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders.” In *Proceedings of the 2019 chi conference on human factors in computing systems,* 1–12.

Choo, J. and Liu, S. 2018. “Visual analytics for explainable deep learning.” *IEEE computer graphics and applications* 38 (4): 84–92.

Chromik, M. and Butz, A. 2021. “Human-xai interaction: A review and design principles for explanation user interfaces.” In *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part II 18,* 619–640. Springer.

Doshi-Velez, F. and Kim, B. 2017. “Towards a rigorous science of interpretable machine learning.” *arXiv preprint arXiv:1702.08608.*

Dry Cactus. 2016. *Poly Bridge.* PC Game, Digital. Dry Cactus.

Dry Cactus. 2020. *Poly Bridge 2.* PC Game, Digital. Dry Cactus.

Frazier, S., Newnan, A., Maheswaran, R., Chang, Y.-H. and Frangoudes, F. 2012. “Team-it: Location-based gaming in real and virtual environments.” In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment,* vol. 8. 1.

Fulton, L. B., Lee, J. Y., Wang, Q., Yuan, Z., Hammer, J. and Perer, A. 2020. “Getting playful with explainable AI: games with a purpose to improve human understanding of AI.” In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems,* 1–8.

Gee, J. P. 2003. “What video games have to teach us about learning and literacy.” *Computers in Entertainment (CIE)* 1 (1): 20–20.

Ghajargar, M., Bardzell, J., Smith-Renner, A. M., Höök, K. and Krogh, P. G. 2022. “Graspable AI: Physical forms as explanation modality for explainable AI.” In *Sixteenth International Conference on Tangible, Embedded, and Embodied Interaction,* 1–4.

GIANTS Software. 2018. *Farming Simulator 19.* PC Game, CD-ROM. GIANTS Software.

Gunning, D. 2017. “Explainable artificial intelligence (xai).” *Defense advanced research projects agency (DARPA), nd Web* 2 (2): 1.

Haque, A. B., Islam, A. N. and Mikalef, P. 2023. “Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research.” *Technological Forecasting and Social Change* 186:122120.

Hodent, C. 2020. *The psychology of video games.* Routledge.

Juul, J. 2013. *The art of failure: An essay on the pain of playing video games.* MIT press.

Koster, R. 2013. *Theory of fun for game design.* ” O’Reilly Media, Inc.”

Liao, Q. V., Gruen, D. and Miller, S. 2020. “Questioning the AI: informing design practices for explainable AI user experiences.” In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems,* 1–15.

Liao, Q. V. and Varshney, K. R. 2021. “Human-centered explainable ai (xai): From algorithms to user experiences.” *arXiv preprint arXiv:2110.10790.*

Liu, S., Wang, X., Liu, M. and Zhu, J. 2017. “Towards better analysis of machine learning models: A visual analytics perspective.” *Visual Informatics* 1 (1): 48–56.

Maxis. 2014. *The Sims 4.* PC Game, CD-ROM. Maxis.

Miller, T. 2019. “Explanation in artificial intelligence: Insights from the social sciences.” *Artificial intelligence* 267:1–38.

Mohseni, S., Zarei, N. and Ragan, E. D. 2021. “A multidisciplinary survey and framework for design and evaluation of explainable AI systems.” *ACM Transactions on Interactive Intelligent Systems (TiiS)* 11 (3-4): 1–45.

Mount Rouke Studios. 2020. *Dr. Derks Mutant Battleground.* PC Game, Digital. Mount Rouke Studios.

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A. and Klein, G. 2019. “Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI.” *arXiv preprint arXiv:1902.01876.*

Myers, C. M., Xie, J. and Zhu, J. 2020. *A Game-Based Approach for Helping Designers Learn Machine Learning Concepts.* eprint: arXiv:2009.05605.

Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S. and Doshi-Velez, F. 2018. *How do humans understand explanations from machine learning systems? An evaluation of the human-interpretability of explanation.* Technical report.

Nguyen, T., Canossa, A. and Zhu, J. 2024. “How Human-Centered Explainable AI Interface Are Designed and Evaluated: A Systematic Survey.” *arXiv preprint arXiv:2403.14496.*

Paradox Interactive. 2020. *Crusader Kings 3.* PC Game, Digital. Paradox Interactive.

Pemberton, D., Lai, Z., Li, L., Shen, S., Wang, J. and Hammer, J. 2019. “AI or Nay-I? Making moral complexity more accessible.” In *Extended Abstracts of the Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts,* 281–286.

Positech Games. 2013. *Democracy 3.* PC Game, Digital. Positech Games.

Ribeiro, M. T., Singh, S. and Guestrin, C. 2016. “”Why Should I Trust You?”: Explaining the Predictions of Any Classifier.” In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data minin.*

Sevastjanova, R., Jentner, W., Sperrle, F., Kehlbeck, R., Bernard, J. and El-Assady, M. 2021. “Questioncomb: A gamification approach for the visual explanation of linguistic phenomena through interactive labeling.” *ACM Transactions on Interactive Intelligent Systems (TiiS)* 11 (3-4): 1–38.

Shrikumar, A., Greenside, P. and Kundaje, A. 2017. “Learning Important Features Through Propagating Activation Differences.” In *Proceedings of the 34th International Conference on Machine Learning.*

Sitzmann, T. 2011. “A meta-analytic examination of the instructional effectiveness of computer-based simulation games.” *Personnel psychology* 64 (2): 489–528.

Spinner, T., Schlegel, U., Schäfer, H. and El-Assady, M. 2019. “explAIner: A visual analytics framework for interactive and explainable machine learning.” *IEEE transactions on visualization and computer graphics* 26 (1): 1064–1074.

Squire, K. and Sasha, B. 2012. “Replaying history: Engaging urban underserved students in learning world history through computer simulation games.” In *Embracing Diversity in the Learning Sciences,* 506–513. Routledge.

Villareale, J. and Zhu, J. 2021. “Understanding mental models of AI through player-AI interaction.” *arXiv preprint arXiv:2103.16168.*

Vlachopoulos, D. and Makri, A. 2017. “The effect of games and simulations on higher education: a systematic literature review.” *International Journal of Educational Technology in Higher Education* 14 (1): 1–33.

Williams, D. 2010. “The mapping principle, and a research framework for virtual worlds.” *Communication Theory* 20 (4): 451–470.

Xie, J., Myers, C. M. and Zhu, J. 2019. “Interactive visualizer to facilitate game designers in understanding machine learning.” In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems,* 1–6.

Zhu, J., Alderfer, K., Furqan, A., Nebolsky, J., Char, B., Smith, B., Villareale, J. and Ontañón, S. 2019. “Programming in game space: how to represent parallel programming concepts in an educational game.” In *Proceedings of the 14th International Conference on the Foundations of Digital Games,* 1–10.

Zhu, J., Liapis, A., Risi, S., Bidarra, R. and Youngblood, G. M. 2018. “Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation.” In *2018 IEEE Conference on Computational Intelligence and Games (CIG),* 1–8. IEEE.

Zhu, J. and Ontañón, S. 2020. “Player-centered AI for automatic game personalization: Open problems.” In *Proceedings of the 15th International Conference on the Foundations of Digital Games,* 1–8.

Zhu, J., Villareale, J., Javvaji, N., Risi, S., Löwe, M., Weigelt, R. and Harteveld, C. 2021. “Player-AI interaction: What neural network games reveal about AI as play.” In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems,* 1–17.